REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

The public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing the burden, to the Department of Defense, Executive Service Directorate (0704-0188). Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to any penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number.

PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ORGANIZATION.

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From - To)*
02-28-2011	Final Report	03/15/2008- 11/30/2010

4. TITLE AND SUBTITLE
Evaluating the Effects of Interface Disruption Using fNIR Spectroscopy

5a. CONTRACT NUMBER
FA9550-08-1-0123

5b. GRANT NUMBER
FA9550-08-1-0123

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)
Robert J.K. Jacob, Leanne M. Hirshfield

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Tufts University, 161 College Ave. Medford, MA 02155

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
AFOSR
875 N. Randolph St.
Suite 325
Arlington, VA 22203

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)
AFRL-OSR-VA-TR-2012-0234

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release

13. SUPPLEMENTARY NOTES

14. ABSTRACT
The primary accomplishment that we achieved during this three year effort was the creation and implementation of a novel usability experiment protocol and a set of machine learning methods that enable us to predict, on the fly, the user state of a given individual. Before we began this research, the majority of brain research, and all fNIRS research could not PREDICT user states. Previous research in non-invasive brain measurement could only predict that two (or more) user states differed from one another. We have had great success publishing our work in part because it offers a large leap forward in the state-of-the-art of non-invasive brain measurement in HCI. We used our techniques to test disruptions that were developed from the DnD project, and we reported on these findings throughout the effort. Building on the techniques and findings from our first 2 ½ years of research, we spent the second half of our final year of funding pursing the measurement of trust and suspicion while users work with computers. We teamed up with a strong group of experts in the trust domain, including interested parties from AFRL at Wright Patterson Air Force Base, where we have visited several times to share, and build on, our research.

15. SUBJECT TERMS
cyber-security, brain measurement, workload, trust, usability testing

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Leanne M. Hirshfield
U	U	U	UU	15	19b. TELEPHONE NUMBER *(Include area code)* 617-314-2801

Reset

Evaluating the Effects of Interface Disruption Using fNIR Spectroscopy
Sponsor No: FA9550-08-1-0123
Final Report

Robert J.K. Jacob/Leanne M. Hirshfield

Tufts University/Hamilton College

1 Introduction

Our AFOSR 6.1 research began as supporting research for the larger AFOSR Denial and Disrupt (D&D) project. The D&D project aims to take an offensive stance toward adversarial computer operators in cyberspace. Cyber-operations techniques have been developed by personnel at Assured Information Security (AIS) to monitor or disrupt an adversarial operator's computer. By introducing subtle disruptions into an adversary's computer system, the D&D project aims to increase user workload, cause distractions, and increase user error rates. Some examples of disruptions are disabling copy and paste functionality, changing the effect of pressing various keyboard keys, causing pop ups to appear on the screen, and slowing down program response time.

By introducing these subtle disruptions into an adversary's computer system we can continually reduce the adversary's threat potential while monitoring his actions. Understanding the effects that sequences of disruptions have on operators is essential to the success of this project. Choosing subtle disruptions and determining how often to disrupt an adversary's computer requires a wealth of experiments and a thorough understanding of human factors. Fundamental questions stemming from the D&D project are:

- How do we measure the effectiveness of various interface disruptions?
- What kinds of disruptions should we choose when:
    - We want to increase subject's error rate without subjects noticing that any disruption is occurring.
    - We want to increase error rate/and or task completion time and we want to frustrate the user without causing suspicion.
- What human performance characteristics do we measure?
- How do we conduct usability experiments to acquire the necessary metrics?

Ideally, usability tests could be conducted to evaluate the effects of the D&D disruptions on computer users. While usability tests can acquire objective, real time metrics of a user's speed and accuracy while (s)he works with an interface, acquiring additional metrics about the user's mental state (such as perceived difficulty, enjoyment, suspicion, boredom, frustration, etc) while working with a computer system involves the use of surveys. These surveys are subjective, and they are often prone to subject biases, and they are administered after a task has been completed, lacking valuable insight into the user's changing experiences while working with an interface.

The novelty of our AFOSR 6.1 research stems from our use of a new, non-invasive brain measurement device to overcome the user state evaluation challenges listed above. We use our non-invasive brain measurements during our usability tests, where we have human users work with a target user interface (this interface could be to an airplane cockpit, an AFRL-sponsored web page, or any software developed for or by the Air Force, to name a few). While users work with the interface on prescribed tasks, we

measure their brain activity, as well as a number of other, more traditional usability metrics. We then use the data from our usability experiment to evaluate the chosen interface design, to clearly specify its strengths (how it supports usability) and its weaknesses (how it limits usability), and to make suggestions about ways to improve (or further impair) the design.

We use electroencephalography (EEG) and functional Near-Infrared spectroscopy (fNIRS) to acquire measures of users' mental states. Unlike other brain devices which require subjects to lie in restricted positions (fMRI), or to drink hazardous materials (PET), EEG and fNIRS can measure users' brain activity in realistic working conditions [1]. This makes EEG and fNIRS appropriate choices for brain measurement in usability testing. Whereas EEG has been available since the early 1900's, fNIRS was not introduced until the 1990's and holds great potential for extremely non-invasive cognitive state measurement. It is significantly easier and faster to set-up on subjects than is EEG, it is less sensitive to noise in the signal, has higher spatial resolution, and is less invasive, allowing for use in real working conditions. Our strong history of research with the promising fNIRS device has placed us at the leading edge of research on non-invasive measurement for usability testing and adaptive system design [2-6].

The 6.1 research that we have conducted over the past three years has enhanced usability testing and the measurement of user states in general, and more specifically, we have applied the findings from our basic research to the usability testing problems stemming from the Denial & Disrupt project.

2 Review of AFOSR Research

At the heart of our previous AFOSR research has been the creation of an experimental protocol that can be used to acquire meaningful measures of users' mental states (i.e., levels of workload, frustration, suspicion) with brain imaging during usability studies [4, 7, 8]. As opposed to most brain measurement research that treats each user state as an independent variable, manipulated in order to control the user's mental state, our protocol enables us to measure each user state as a dependent variable during the usability study (as the point of a usability study is to determine the *unknown* effects that an interface has on computer user). As a reminder, the general protocol is as follows:

1) Researchers gather benchmark tasks from cognitive psychology that elicit high and low levels of brain activity on a range of target cognitive resource(s) such as visual search, working memory, response inhibition, and tasks that elicit emotional responses with set levels of valence and arousal. These exercises are henceforth referred to as *cognitive benchmark tasks.*

2) Researchers create a set of tasks that users will complete with the UI to be evaluated. These tasks are referred to as *UI tasks.*

3) An experiment is run where users complete the *cognitive benchmark tasks*, yielding a measure of their brain activity while experiencing high and low brain activity levels in their various cognitive subsystems. This is called the *training* period.

4) Users also work with the *UI tasks*. This is called the *testing* period. Brain activity is measured throughout the experiment.

5) Data analysis tools are used to find similarities and differences in the users' brain activity while working with the tasks in the *training* and *testing period*.

With this protocol, researchers can make connections between the brain activity experienced by users' low level cognitive resources while they work with complex user interfaces.

2.1 Year One Review

We focused on the measurement of mental workload in controlled laboratory settings during our first year of research. As described in detail in our year one annual report, our first year of research focused on:

- Measuring users' mental workload with fNIRS in tightly controlled experiments[5, 9, 10].
- The creation of an initial suite of analysis algorithms for use with fNIRS data.
- Combining EEG and fNIRS for measuring a wider range of users' mental workload states[3]
- Developing the novel protocol (described above) to measure workload as a dependent variable, when the workload caused by a given interface is complex and difficult to interpret.
- The conducting of a pilot experiment with AIS and simple interface disruptions to test out our analysis techniques and novel protocol for the evaluation of these interface disruptions [8].

Our second year of research built on the first year of research. In the second year of research our lead researcher, Dr. Leanne Hirshfield, completed her PhD at Tufts and accepted a research position at Hamilton College. The second year of research, and our ongoing research in year three, reflects a joint research effort by students and professors at Hamilton College and Tufts University.

2.2 Year Two Review

During our second year of research we iterated on the novel protocol described above and in [4], with the goal of validating the protocol. We also worked to extend the protocol to evaluate a variety of interface designs with the measurement of several mental states. In particular, during our second year of funding we made five notable research accomplishments, which were discussed in detail in our year two annual report:

1) We ran an experiment using interface disruptions from the Denial and Disrupt (D&D) Project and we measured the effects that two disruptions had on computer users' level of working memory load [8].
2) We began to develop new fNIRS feature extraction and machine learning techniques that are more powerful than those developed in year 1 (see publications on SAX and Dynamic Time Warping Techniqus) [4].
3) We ran an experiment to further validate our experiment protocol in a realistic usability experiment. This experiment focused on evaluating two generic user interfaces (rather than evaluating interface variations from the D&D project). The experiment demonstrates the generality of our experimental protocol and analysis techniques.
4) We acquired a custom fNIRS probe and adaptive filtering algorithms to remove noise in the fNIRS signal caused by subject movement; thus enabling us to run realistic usability studies where subjects can move freely [8].

2.3 Year Three Review

In our final year of funding we tied up several loose ends in our research and we began preliminary experiments for a new avenue of research. Our work included finishing the iterations in development on our machine learning techniques, validating the robustness of our new techniques on a range of datasets, building a new lab with the help of DURIP funding, and beginning research to measure the user states of trust and suspicion while users work with their computers.

2.3.1 Year Three Review: New Machine Learning Techniques

As described above, an important aspect of our research has been the development and validation of the usability experiment protocol that can be used to measure user states on the fly. The human brain is extremely complex, and making the leap from measurement of brain activity to a meaningful user state is non-trivial. Our protocol enables us to make these complex jumps. As a reminder, the usability experiment protocol is as follows:

1) Researchers gather benchmark tasks from cognitive psychology that elicit high and low-levels of WL on a range of target cognitive resource(s) such as visual search, WM, and response inhibition. We refer to these exercises as *cognitive benchmark tasks.*

2) Researchers create a set of tasks that users will complete with the UI to be evaluated. We refer to these as *UI tasks.*

3) An experiment is run where users complete the *cognitive benchmark tasks*, yielding a measure of their brain activity while experiencing high and low WL levels in their various cognitive subsystems. Users also work with the *UI tasks.* Brain activity is measured throughout the experiment.

4) fNIRS data from the *cognitive benchmark tasks* are used as training data to build a machine learning classifier. The fNIRS data from the *UI tasks* is input into the machine learning classifier as testing data for the classifier. The classifier outputs the level of cognitive load experienced by each user while working with a given *UI task*.

Step 4), above was the focus of much of our third year of research. We finished the development of algorithms to use the fNIRS data from our usability experiments to predict the user state of the person working with the computer during the experiment.

The techniques we finished developing in the third year provide significantly higher classification accuracies than the techniques we used in years 1 (neural nets and timeseries data) and 2 (time series data and SAX or Dynamic Time Warping with K-nearest neighbor classifier) of the research. Before building our machine learning classifier, we preprocess the fNIRS data to remove noise from heartbeat, motion artifacts, and breathing. We then extract the following features from our preprocessed fNIRS data: *largest value, smallest value, average, slope, time to peak*, and *full width at half maximum.* We extract these features for the first and second half of each task.

For each participant, we input training data from our benchmark tasks to build a Naïve Bayes classifier. This classifier slightly outperforms other classifiers that we attempted such as Support Vector Machines, Neural Nets, and K-nearest neighbor classifiers. For example, Figure 1 depicts our process for creating a

Naïve Bayes classifier to predict the WM load (high or low) associated with a given UI task. First, the data from each of the low and high benchmark WM tasks was used as training data to build a Naïve Bayes Classifier. This results in a WM classifier that predicts whether or not a given test instance has a high or low WM load.

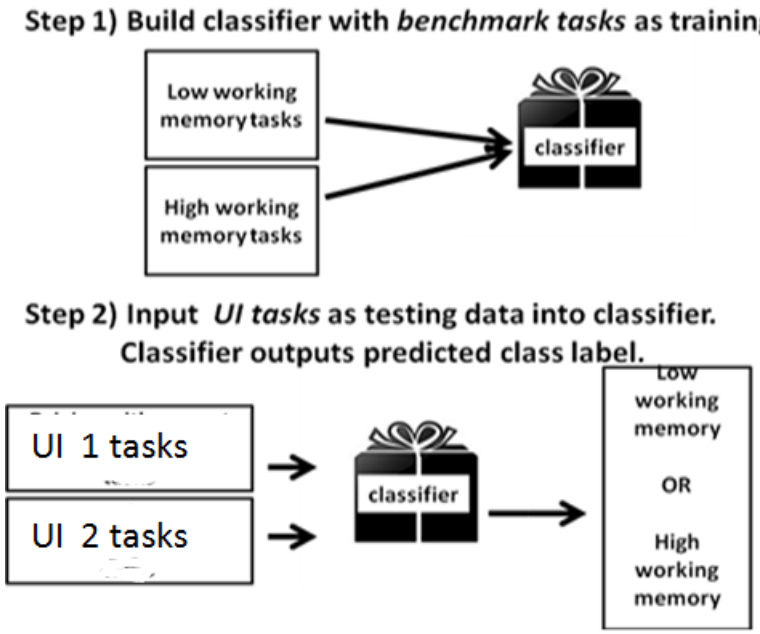

Figure 1: The *WM classifier* is built using WM benchmark tasks. The UI tasks are fed into the classifier as test instances.

We've had promising levels of accuracy using our new machine learning techniques on our usability experiment data. We ran the new technique on all old fNIRS datasets and achieved an average increase in accuracy of 15%! We'll be presenting these machine learning techniques at the premiere conference in Human Computer Interaction this April. At the conference, we'll present our machine learning classifier's success at predicting the level of load placed on participants' visual search, working memory, and response inhibition cognitive resources while working with a simulated driving task and a web search task [7]. Our ability to predict load placed on such fine tuned cognitive resources as these shows great potential for using the protocol and machine learning methods to measure a range of user states such as workload, frustration, engagement, trust, and suspicion.

Thus, and we re-iterate this point to ensure it is clear**. One of the strongest accomplishments that we reached during this three year effort was the creation, implementation, and validation of a novel usability experiment protocol and a set of machine learning methods that enable us to predict, on the fly, the user state of a given individual.** Before we began this research, the **vast** majority of brain research (with the exception of some of the work by Berka et al [11]), and **all** fNIRS research could not PREDICT user states. This research could only predict that two (or more) user states differed from one another. We have had great success publishing our work in part because it offers a large leap forward in the state-of-the-art of non-invasive brain measurement in HCI.

2.3.2 Year Three Review: Opening of New Lab

During our last year of funding we were also recipients of a $468,000 DURIP award from AFOSR. With the equipment funding we created a cutting edge usability lab with a fully synchronized instrument for acquiring concurrent measurements of cognitive, physiological, and behavioral user state metrics. With an eye toward their use in realistic human-computer interactions, our non-invasive equipment includes a wireless, lightweight Electroencephalograph (EEG), a functional near infrared spectroscopy (fNIRS) device, an eye tracker that can be used both inside and outside (most eye-trackers are sensitive to light and must be used in controlled light settings), and a wireless device for measuring galvanic skin response (GSR). With our usability software, we also have the capability to measure more traditional metrics such as speed, accuracy, key presses, mouse movement, and screen captures. By acquiring concurrent recordings of cognitive, physiological, and behavioral user metrics during our usability studies, we are able to measure a person's user state in real time, and with a high degree of accuracy throughout our human subject studies.

As described previously, a key element of our research is the use of non-invasive brain measurement to provide real time, objective information about users' mental states. In particular, expertise with fNIRS places us at the cutting edge of our research domain. The few fNIRS devices that are being used in the research domain cost less than $50k, and they are limited by the number of channel measurements (usually 16) available, and by the fact that the measurement sensors are usually limited to placement on subjects' foreheads. With the DURIP funding, our fNIRS device is Hitachi Medical's ETG-4000. The $315k piece of equipment was selected because of its wide range of probes and holders that enable a variety of assessments. This system can be configured to simultaneously take 24, 48 or 52 channel measurements, and it can be used to target various regions of the brain cortex (as opposed to just the prefrontal cortex). With this device, our usability work will be at the cutting edge of non-invasive brain measurement research.

2.3.2 Year Three Review: Measuring Suspicion

During our third year we also began research to quantify the user states of 'trust', 'distrust', and 'suspicion', that occur when a user works with a computer system. On a high level, our goal was to measure the cognitive (i.e. the fNIRS and EEG brain data), physiological (i.e., galvanic skin reponse, heartbeat, respiration), and behavioral metrics (i.e., keypresses, mouse movement, speed, accuracy) that occur when a user loses trust and becomes suspicious that his computer system has been hacked. The two primary AFOSR related end goals that stem from this basic research are:

1) ***Monitoring of Adversarial Computer Operators:*** It is worth noting that many of the behavioral metrics that we will measure can be measured remotely; meaning that we can simply record information on a user's computer, as opposed to requiring the user to wear some sort of measurement equipment. Currently, these remote metrics can be acquired in our usability experiments, and we can make controlled connections between the remote metrics and our concurrent recordings of cognitive and physiological data. In the future we could acquire these remote metrics directly from the computer of an adversarial computer operator. Therefore, by making ties to the remote metrics that occur while users are experiencing changes in their level of trust toward the computer, we can monitor adversarial computer operators and determine their

level of trust toward their computer system—which we have hacked into (i.e., are they suspicious of our hack?). We can incorporate trust repair tactics as needed if we determine that the adversary has detected our security breach. This research has direct implications for the AFOSR Denial & Disrupt project, where we can run usability studies to determine the threshold at which certain interface disruptions become too overt, causing users to become suspicious of a computer hack.

2) ***Training Military Personnel***: By understanding the cognitive and emotional aspects of the trust relationship between a trainee and his computer system, we can better train our military personnel to eliminate cognitive shortcuts that sacrifice computer security in return for lower cognitive load and lower stress. Research suggests that people tend to be 'cognitive misers'. We often make cognitive shortcuts in order to minimize mental workload. One of these cognitive shortcuts includes adapting a *truth bias*, where people tend to believe that all people they interact with are being honest and truthful. The *truth bias* is a cognitive shortcut that keeps people from becoming suspicious of other people, as suspicion causes people to critically think about their interactions, causing a high level of cognitive load. We hypothesize that many personnel in the US military adapt this *truth bias* while working with their computer systems; it's simply too much effort for them to think critically about the validity of the information being presented in their interactions with their computer. Instead, they assume that they are usually not being hacked into, and they may miss a vital security breach. In the future, non-invasive cognitive, behavioral, and physiological measurement devices could be used in training sessions to ensure that military personnel develop the correct cognitive and physiological mental state to accurately detect security breaches.

In our final year of funding we joined together with a team of experts in the trust and management domains. Our team of trust collaborators included Lt. Col Alex Barelka from AFRL at Wright Patterson AFB, Dr. Roger Mayer from NC State University, and Dr. Philip Bobko from Gettysburg College. We worked (and continue to work) to create new models of trust, distrust, and suspicion that describe the interactions between a person and that person's computer system. We also created plans to empirically validate our models and to measure a user's changing level of trust while working with his or her computer system.

Our goal within this larger team effort was to find the cognitive, physiological, and behavioral correlates that predict users' level of trust, distrust, and suspicion toward their computer system.

After conducting a thorough literature review and collaborating with many experts in the trust domain, we were encouraged to find that this research is novel for several reasons (which is why we have applied for follow-on funding):

1) Most research doesn't look at the changing relationship from trust -> distrust over time. They usually place someone in a low trust state (i.e. "The person you are about to interact with might be lying to you."). We'll be exploring the nature of the changing trust relationship over time, and we'll be looking at the effects of repair tactics on the trust relationship.
2) To the best of our knowledge, there is little to no research exploring the cognitive, physiological, and behavioral changes that occur in an individual who is experiencing changing levels of trust <-> suspicion <-> low trust. The majority of the literature looks at the cognitive, physiological, and

behavioral changes that occur in a person who is being deceitful, not at those metrics occurring when a person is detecting deceitfulness.

3) There is very little research available that looks at the trust relationship between a person and that person's networked computer system**.**
4) Being able to predict a user's level of trust with unbiased, real time data while he or she works with a networked computer can have far reaching applications for the military.

In our final year of funding we ran a set of preliminary experiments to determine the feasibility of manipulating, and measuring, trust and suspicion during human-computer interactions. Our long term goal is to measure changing levels of trust during human-computer interactions. As a first step, we ran preliminary experiments that isolate the cognitive components of workload, frustration, and surprise [12].

Preliminary Trust Experiment

We conducted an experiment that aimed to discover relationships between a computer user's level of trust and that user's changing cognitive and emotional states. To do so we asked participants to sit in front of a standard-size computer monitor and interact with a computer console on the computer screen. They then played a version of the "Trust Game," developed by Berg [13], which has been used in many experiments dealing with trust, risk-taking, and money management [14]. In our version of the "Trust Game" both the computer and the user began with a fictional $10. The user and computer would take turns sending some amount of money, ranging from $0-$10, back and forth to one another. Each time some amount of money was sent between the user and computer, the amount sent was tripled while en route. In an ideal, high-trust scenario, the computer and the user would always send a the maximum amount of money back and forth to one another—maximizing the gain possible for each of the two Trust Game 'players'. This process was repeated 23 times.

For the first eight transactions, the computer acted "trustworthy" in the sense that it typically returned a high amount of money back to the participant. For the next nine transactions, the computer acted with a mix of trustworthy and untrustworthy behavior, sometimes returning a high amount, and other times returning a low amount of money back to the participant. For the last six transactions the computer acted in a wholly untrustworthy manner, regularly returning a very low amount of money back to the participant.

We used data from Self Assessment Manikins which were administered after every six transactions to gauge the user's cognitive and emotional state. Additionally, we collected the amount given and percentage returned for each transaction, as well as information regarding the subjects' self-rated locus of control, trust, and computer familiarity.

From this information we gleaned that the median amount given increased during the computer's trustworthy state, varied highly during the computer's erratic state, and decreased during the computer's untrustworthy state. Furthermore, we analyzed the data from the Self Assessment Manikins, turning users self reported measures of valence, arousal, and dominance into a set of discrete user states. We found a direct correlation between the trustworthiness of the computer agent and the user's reported measures of workload, frustration, and surprise. Our results showed that overall workload increased throughout the experiment, as did frustration and surprise.

We were also interested in whether relationships existed between frustration, surprise, and workload. We found that significant correlations between frustration and workload in all four surveys existed. We also found significant correlations between workload and surprise in surveys 1, 3, and 4, as shown in Table 1 below.

	Survey 1	Survey 2	Survey 3	Survey 4
Frustration & Workload	$r(27)= .565$ $p<.01$	$r(25)=.733$ $p<.01$	$r(25)=.67$ $p<.01$	$r(26)=.739$ $p<.01$
Workload and Surprise	$r(27)=.559$ $p<.01$	$r(25)=.391$ $p<.06$	$r(25)=.638$ $p<.01$	$r(26)=.478$ $p<.02$

Table 1: Significant Correlations between workload/frustration, and workload/surprise

Our results suggested that as computer users lost trust with the simulated agent they were interacting with in the Trust Game, the user states of workload, surprise, and frustration were directly correlated with the users' changing level of trust.

Linking Trust in Human-Computer Interactions to Surprise, Workload, and Frustration

While these results are restricted to the version of the Trust Game that users played, we hypothesized that the same user states will influence the level of trust in more realistic interactions between users and their computer systems. As an illustrative example, consider John Doe's interaction with his computer over the course of a year:

When John's computer was functioning properly, he had high trust in his interactions with and through his computer. During these high trust times, John had low levels of frustration, workload, and surprise—all interactions with his computer seemed to proceed as expected. However, one day John visited a new website and suddenly he noticed hundreds of pop ups infiltrating his screen (i.e., surprise and frustration). He later found out that his computer had a virus, which likely came from the site with all of the pop ups (frustration). Later on that year, , John was Instant Messaging with his friend Alice. Alice was a classmate of John's and a user with her name as a username contacted John via IM. After a few minutes of messaging with who he presumed to be Alice, John began to become wary of the interactions. The IMer was not writing in a way that was consistent with Alice. John began to interact very cautiously with the IMer, hoping to determine if the person was indeed an imposter (workload, frustration, surprise). Also, over the course of time, John's computer became very slow because he downloaded too many programs and add-ons. He was frustrated while using it because it took a long time for him to get things done and he found it difficult to keep focused on the task at hand while waiting long intervals for his computer to catch up to his train of thought (frustration and workload).

All of these occurrences caused John's level of trust during his computer interactions to be lowered. We hypothesize that we can use measures of users' workload, frustration, and surprise to indicate that users' level of trust. In the next sections we describe the user states of workload, frustration, and surprise, and we describe our initial research attempts in year 3 to measure these user states objectively.

Measurement of Surprise, Workload, and Frustration

Acquiring quantitative data about computer users is a continual challenge for researchers in HCI. Although we can accurately measure task completion time and accuracy, measuring factors such as mental workload, frustration, and distraction are often done by qualitatively observing users or by administering subjective surveys to users. These surveys are often taken after the completion of a task, potentially missing valuable insight into the user's changing experiences throughout the task. They also fail to capture internal details of the operator's mental state. To address these evaluation issues, much current research focuses on developing objective techniques to measure, in real time, user states such as workload, frustration, and surprise [15-17]. Although this ongoing research has advanced user experience measurements in the HCI field, finding accurate and non-invasive tools to measure computer users' states in real working conditions remains a challenge. The user states that are addressed by this research are the states of workload, frustration, and surprise.

Surprise

Surprise has previously been measured in HCI studies using facial analysis software [18] as well as using Skin Conductivity, blood volume and heart rate [19]. Detecting surprise with electroencephalography (EEG) is a topic of much research in Psychophysiology. Surprise can be indicated by the presence of an Error-Related Potential (ErrP), in which EEG data contains error-related negativity (a sharp negative deflection around 80ms)[20], often followed by error-related positivity (a slow positive wave)[21]. ErrPs can be found both when a user makes an error and when a user notices an error made by the machine [22]. This makes the measurement of ErrPs useful in HCI-based interface analysis.

Frustration

Frustration is an important metric in HCI. Lazar, et al. studied frustration with computer interfaces by having employees that used computers in their workday keep journals that tracked their ongoing frustration as they used their routine computer programs [23]. The results showed that word processing and email were reported as the most frustrating activities, and that participants wasted an average of forty percent of their time trying to solve unnecessarily frustrating problems.

Biological methods of measuring frustration allow researchers to collect more reliable data. Scheier, et al., induced frustration in users with a mouse that sporadically froze and inhibited the user from winning a game [24]. A Hidden Markov Model analyzed skin conductivity, blood volume pressure and the state of the mouse, eventually learning the manifestations of frustration. The results indicate that a user's affective state can be automatically discriminated from events in their physiology [24]. Most recently, BCI devices enable automated detection of user frustration through discovering patterns in brain activity. Reuderink et al. developed an affective version of Pacman which places the user in a state of frustration [16]. They found significant differences in EEG activity during periods of frustration and the normal state.

Workload

The ability to acquire objective, real-time measures of a computer user's mental workload while (s)he works with a computer would be valuable to the field of HCI. Adaptive interfaces could adapt in real-time to a given user based on his or her current level of workload, keeping that user in *the flow* [25]. Also, measures of users' mental workload could be acquired during usability studies to help interface designers to pinpoint areas of the interface that may be un-intuitive for users [4, 26].Researchers have successfully

used EEG or fNIRS to measure elements of mental workload such as working memory[7, 9, 27, 28], response inhibition [7, 29], visual search [7, 30], as well as a myriad of other executive processes [31, 32].

Preliminary Trust Experiment: Measurement Oriented Experiments

We conducted preliminary experiments during our third year of funding where we attempted to manipulate and measure the user states of surprise, frustration, and workload. In the future, we aim to acquire real time measures of these user states in order to predict one's level of trust during his or her computer interactions.

Surprise Experiment

An important component of trust is the moment of surprise; that is the moment when a person notices that something 'unexpected' has occurred in the computer system. This could be the moment users notice that a virus is on their computer, or the moment they realize that the person they are IMing with may be an imposter. To measure this, we exploited the oddball paradigm in order to elicit surprise. Three participants completed an experiment that was created using Eprime in which they pressed two different buttons depending on the position of an oval on the screen. The oval was in one of two positions, located either on the far left or the far right side of the screen. When the oval was on the left side of the screen the subjects were instructed to press the 'z' button, and when the oval was on the right side of the screen they were instructed to press the 'm' button.

Immediately following the subject response a feedback screen indicated whether or not the subject had pressed the correct key. Subjects completed 150 tasks where they simply hit the 'z' or 'm' keys to indicate the position of the oval on the screen. During the first 20 tasks, the feedback for the subjects was as expected. During the last 130 tasks, we randomly selected 15% of the tasks to provide incorrect, , *or surprising feedback* to the user. In other words, 15% of the time, when subjects pressed the 'z' key, the feedback indicated that the 'm' key had been pressed, and vice versa.

The EEG used in the study was Advanced Brain Monitoring's b-alert wireless 10 channel EEG. Data was sampled at 256Hz (www.b-alert.com). The non-invasive EEG is an ideal brain monitoring device for use in human-computer interaction studies, where it may be important to keep participants comfortable while completing tasks in realistic working conditions.

The Eprime software sent markers to the EEG immediately before the subject saw the feedback screen. In this way, we planned to search for the presence of an ErrP that was caused when the surprising feedback occurred during 15% of the tasks.

Data Analysis and Results of Surprise Experiment

We used a similar procedure as Ferrez et al. [22] to preprocess our EEG data for classification. We took the data from the moment the feedback occurred through to 650ms after the feedback was shown for channels Cz and Fz. Like Ferrez et al., we chose these channels because ErrPs are usually found in a fronto-central distribution along the midline [22] Each temporal section of data was associated with one of two class labels: control or surprise, indicating whether or not the feedback the subject saw at that moment was the expected feedback or the surprising feedback.. We applied a 1-10 Hz bandpass filter as ErrPs have a relatively slow cortical potential. We downsampled our data from 256Hz to 128Hz and input our resulting timeseries data into a weighted K-nearest neighbor classifier (k = 3) with a Dynamic Time

Warping distance measure. We ran our classification separately for each subject. Results are in Table 2. We were able to distinguish between the control and surprising feedback conditions with an average of 71% accuracy for our three subjects.

Table 2: Classifier accuracy distinguishing between the control and surprising feedback.

	sub1	sub2	sub3	average
Classifier Accuracy	70%	74%	68%	71%

Frustration Experiment

During the frustration experiment six subjects completed a series of nback tasks [7, 27], which have been used in many experiments to manipulate working memory. In the 1-back task, depicted in Figure 2, subjects must indicate whether the current letter on their computer screen is a match ('m'), or not a match ('n') to the letter that was shown 1 screen previously.

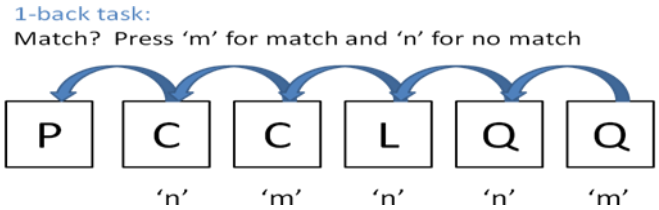

Figure 2: Depiction of the 1 back task.

Each task lasted 30 seconds with a rest time of 20 seconds between tasks. Half of the 1back tasks were completed by subjects as expected. However, during the other half of the 1back tasks, internet pop ups such as the one shown in Figure 3, were introduced into the computer systems. Subjects were told to finish the nback tasks as quickly as possible and with the highest accuracy possible. Six subjects (3 female, 3 male) completed the experiment. Subjects were all Tufts undergraduate students. A randomized block design with eight trials was used in this experiment.

Figure 3: An example of a pop up in the frustration experiment.

In this experiment we used an OxyplexTS (ISS Inc. Champagne, IL) frequency-domain tissue spectrometer with two optical probes. Each probe has a detector and four light sources. Each light source emits near infrared light at two separate wavelengths (690nm and 830nm) which are pulsed intermittently in time. This results in 2 probes x 4 light sources x 2 wavelengths = 16 light readings at each timepoint (sampled at 6.25Hz).

Data Analysis and Results of Frustration Experiment

All subjects were interviewed following the experiment. All subjects indicated that the pop ups were a source of frustration throughout the experiment. We computed all machine learning analyses separately for each subject. For each subject, we recorded 16 channel readings throughout the experiment where we

refer to the readings of one source detector pair at one wavelength, as one *channel*. We normalized the intensity data in each channel by their own baseline values. We then applied a moving average band pass filter to each channel (with values of .1 and .01 Hz) and we use the modified Beer-Lambert Law[12] to convert our light intensity data to measures of the relative changes in oxygenated (HbO) and deoxygenated hemoglobin (Hb) concentrations in the brain. This resulted in eight readings of HbO and eight readings of Hb data at each timepoint in the experiment. We then averaged together the channels from the left side of the head and the channels on the right side of the head, giving us 4 time series for each subject; 1) HbO on the left side of the head, 2) HbO on the right side of the head, 3) Hb on the left side of the head, and 4) Hb on the right side of the head. We then input these time series into a weighted KNN classifier (k = 3) with a distance measure computed via Symbolic Aggregate Approximation (SAX). For more information on SAX, see [33]. As shown in Table 3, we were able to distinguish between the control 1back tasks and the frustrating 1back tasks with an average of 73% accuracy across the six subjects.

Table 3: Classifier accuracy at distinguishing between the control (1back) and frustrating (1back with pop-ups) conditions.

	sub1	sub2	sub3	sub4	sub5	sub6	average
Classifier Accuracy	69%	81%	63%	75%	75%	75%	73%

Workload Experiments

We have conducted several experiments, using the fNIRs device described above, to measure various aspects of mental workload. Using this device we have:

1) Used machine learning techniques to classify, on a single trial basis, the load placed on users visual search, working memory, and response inhibition resources [7].
2) Used machine learning techniques to classify various levels of working memory load in a simple counting and addition task [5].
3) Used machine learning techniques to distinguish between spatial and verbal working memory[4].

Preliminary Trust Experiment Conclusion

While there is certainly more work to be done, the preliminary experiments and research that we conducted during our third year of the project show promise for linking the user states of workload, surprise, and frustration to trust and suspicion. We are continually working with our trust colleagues to develop high level models of trust and suspicion, and to measure these user states using the suite of devices in our new lab.

3 Conclusion

In conclusion, we had a very enjoyable, and very fruitful three years while working on this AFOSR effort. One of the greatest accomplishments that we reached during this three year effort was the creation and implementation of a novel usability experiment protocol and a set of machine learning methods that enable us to predict, on the fly, the user state of a given individual. Before we began this research, the **vast** majority of brain research (with the exception of some of the work by Berka et al), and **all** fNIRS

research could not PREDICT user states. This research could only predict that two (or more) user states differed from one another. We have had great success publishing our work in part because it offers a large leap forward in the state-of-the-art of non-invasive brain measurement in HCI. We used our techniques to test disruptions that were developed from the DnD project, and we reported on these findings throughout the effort. Building on the techniques and findings from our first 2 ½ years of research, we spent the second half of our final year of funding pursing the measurement of trust and suspicion while users work with computers. We teamed up with a strong group of experts in the trust domain, and we began preliminary research with this group. With members at Wright Patterson and at AFRL showing interest in this new avenue of research, we look forward to continuing this work in the future.

4 References:

1. Izzetoglu, M., et al., *Functional Near-Infrared Neuroimaging.* IEEE Trans Neural Syst Rehabil Eng, 2005. **13**(2): p. 153-9.
2. Hirshfield, L.M., *Enhancing Usability Testing with Functional Near Infrared Spectroscopy*, in *Computer Science*. 2009, Tufts University: Medford, MA.
3. Hirshfield, L.M., et al. *Combining Electroencephalograph and Near Infrared Spectroscopy to Explore Users' Mental Workload States*. in *HCI International*. 2009: Springer.
4. Hirshfield, L.M., et al. *Brain Measurement for Usability Testing and Adaptive Interfaces: An Example of Uncovering Syntactic Workload in the Brain Using Functional Near Infrared Spectroscopy*. in *Conference on Human Factors in Computing Systems: Proceeding of the twenty-seventh annual SIGCHI conference on Human factors in computing systems*. 2009.
5. Hirshfield, L.M., et al. *Human-Computer Interaction and Brain Measurement Using Functional Near-Infrared Spectroscopy*. in *Symposium on User Interface Software and Technology: Poster Paper*. 2007: ACM Press.
6. Girouard, A., et al., *From Brain Signals to Adaptive Interfaces: using fNIRS in HCI*, in *(B+H)CI: The Brain in Human-Computer Interaction and the Human in Brain-Computer Interfaces*, D.S. Tan and A. Nijholt, Editors. 2010, Springer.
7. Hirshfield, L., et al. *This is your brain on interfaces: enhancing usability testing with functional near infrared spectroscopy*. in *SIGCHI*. 2011 (in press): ACM.
8. Hirshfield, L.M., *Enhancing Usabiltiy Testing with Functional Near Infrared Spectroscopy*, in *Computer Science*. 2009, Tufts University: Medford, MA.
9. Sassaroli, A., et al., *Discrimination of mental workload levels in human subjects with functional near-infrared spectroscopy.* accepted in the Journal of Innovative Optical Health Sciences, 2009.
10. Girouard, A., et al. *Distinguishing Difficulty Levels with Non-invasive Brain Activity Measurements*. in *Proc. INTERACT Conference*. 2009.
11. Berka, C. and D. Levendowski, *EEG Correlates of Task Engagement and Mental Workload in Vigilance, Learning and Memory Tasks.* Aviation Space and Environmental Medicine, 2007. **78**(5): p. B231-B244.
12. Hirshfield, L., et al. *Trust in Human-Computer Interactions as Reflected by Workload, Frustration, and Surprise*. in *HCI International 2011 14th International Conference on Human-Computer Interaction*. 2011 (in press): Springer.
13. Berg.J., J. Dickhaut, and K. McCabe, *Trust, Reciprocity, and Social History.* Games and Economic Behavior, 1995. **10**: p. 122-142.

14. Lewicki, R., et al., *Trust and Distrust: New Relationships and Realities.* The Academy of Management Review, 1998. **23** (3): p. 438-458.
15. Mandryk, R., M. Atkins, and K. Inkpen, *A continuous and objective evaluation of emotional experience with interactive play environments*, in *Proceedings of the SIGCHI conference on Human factors in computing systems*. 2006, ACM Press: Canada.
16. Reuderink, B., A. Nijholt, and M. Poel, *Affective Pacman: A Frustrating Game for Brain-Computer Interface Experiments.* Intelligent Technologies for Interactive Entertainment, Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering, 2009.
17. Savran, A., et al. *Emotion Detection in the Loop from Brain Signals and Facial Images*. in *eNTERFACE'06*. 2006. Dubrovnik, Croatia.
18. Ward, R., *An analysis of facial movement tracking in ordinary human-computer interaction.* Physiological Computing, 2004. **16**(5): p. 879-89.
19. Ward, R. and P. Marsden, *Physiological responses to different web page designs.* International Journal of Human Computer Studies, 2003. **59**: p. 199–212.
20. Chavarriaga, R., P. Ferrez, and J. Millán, *To Err is Human: Learning from Error Potentials in Brain-Computer Interfaces.* ADVANCES IN COGNITIVE NEURODYNAMICS, 2008: p. 777-782.
21. Nieuwenhuis S, et al., Psychophysiology, Error-related brain potentials are differentially related to awareness of response errors: Evidence from an antisaccade task: p. 752–760.
22. Ferrez, P. and J. Millán. *You Are Wrong!---Automatic Detection of Interaction Errors from Brain Waves*. in *Proceedings of the 19th International Joint Conference on Artificial Intelligence*. 2005.
23. Lazar, J. and A.S. Jones, *Workplace user frustration with computers: an exploratory investigation of the causes and severity.* Behaviour & Information Technology, 2006: p. 239-251.
24. Scheirer, J., et al., *Frustrating the user on purpose: a step toward building an affective computer.* Interacting with Computers, 2002: p. 93-118.
25. Csikszentmihalyi, M., *Flow: The Psychology of Optimal Experience.* 1991.: Harper Collins. 320.
26. Lee, J.C. and D.S. Tan, *Using a low-cost electroencephalograph for task classification in HCI research*, in *Proceedings of the 19th annual ACM symposium on User interface software and technology*. 2006, ACM Press: Montreux, Switzerland.
27. Grimes, D., et al. *Feasibility and Pragmatics of Classifying Working Memory Load with an Electroencephalograph*. in *CHI 2008 Conference on Human Factors in Computing Systems*. 2008. Florence, Italy.
28. Gevins, A., et al., *High-Resolution EEG Mapping of Cortical Activation Related to Working Memory: Effects of Task Difficulty, Type of Processing, and Practice.* Cerebral Cortex, 1997.
29. Schroeter, M.L., et al., *Near-Infrared Spectroscopy Can Detect Brain Activity During a Color-Word Matching Stroop Task in an Event-Related Design.* Human Brain Mapping, 2002. **17**(1): p. 61-71.
30. Anderson, E.J., et al., *Involvement of prefrontal cortex in visual search.* Experimental Brain Research, 2007. **180**(2): p. 289-302.
31. Tanida, M., et al., *Relation between asymmetry of prefrontal cortex activities and the autonomic nervous system during a mental arithmetic task: near infrared spectroscopy study.* Neuroscience Letters, 2004. **369**(1): p. 69-74.
32. Joanette, Y., et al., *Neuroimaging investigation of executive functions: evidence from fNIRS.* PSICO, 2008. **39**(3).
33. Lin, J., et al. *A Symbolic Representation of Time Series, with Implications for Streaming Algorithms.* in *In proceedings of the 8th ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery*. 2003. San Diego, CA.